

Smoothed L-Estimation of Regression Function*

Julien Tamine

Université de la Méditerranée
GREQAM Vieille Charité
2 rue de la Charité
13002 Marseille
France

Pavel Čížek

Center for Applied Statistics and Economics
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
Spandauer Str.1, D-10178 Berlin
Germany

Wolfgang Härdle

Center for Applied Statistics and Economics
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
Spandauer Str.1, D-10178 Berlin
Germany

Abstract

The Nadaraya-Watson estimator of regression is known to be highly sensitive to the presence of outliers in the sample. A possible way of robustification consists in using local L-estimates of regression. Whereas the local L-estimation is traditionally done using an empirical conditional distribution function, we propose to use instead a smoothed conditional distribution function. We show that this smoothed L-estimation approach provides computational as well as statistical finite sample improvements. The asymptotic distribution of the estimator is derived under mild β -mixing conditions.

Keywords: nonparametric regression, L-estimation, smoothed cumulative distribution function

*The financial support of the German Science Foundation (grant SFB 373, A1), is gratefully acknowledged. The first author was also supported by Greqam, Université de la Méditerranée.

1 Introduction

The nonparametric estimation of regression functions has received much attention in the literature and one of the most widely used estimator is, without any doubt, the Nadaraya-Watson estimator by Nadaraya (1964) and Watson (1964). This estimator, being a local average of the response variable, is highly sensitive to the presence of outliers in the data; see Barnett and Lewis (1979) for a general discussion of the concept of an outlier. Indeed, possible outliers do not only increase the variance of the estimator, but can also create fictitious peaks and therefore structure in the estimation. In order to robustify this estimator, Boente and Fraiman (1994) proposed to use a local L-estimate such as local α trimmed means instead of a locally weighted average. Their procedure consists in using an empirical conditional distribution function that allows for the estimation of the amount of data to be discarded. We demonstrate that the choice of an empirical conditional distribution is not appropriate and our aim in this paper is to show that the use of a smoothed conditional distribution function has substantial advantages. First, it does not need the computation of estimates of the local conditional cumulative distribution function at every point of the sample but only on an integration grid. As a consequence, it will be shown to be less computationally intensive. Second, following theoretical arguments of Fernholz (1997) for non-conditional L-estimates, we expect our estimator to have better finite sample properties.

Our work is organized as follows: in Section 2, we describe both the empirical and the smoothed estimators of conditional distribution function and we explain how they can be used to estimate the conditional L-estimator of regression. In Section 3, we give asymptotic bounds for the smoothed conditional distribution function and we derive the asymptotic distribution of the smoothed L-estimator. In Section 4, we

show why the smoothed estimator can be computationally less time consuming than the empirical one and we present the results of a simulation that illustrates these improvements. In Section 5, we briefly recall the arguments given by Fernholz in favor of smoothing (in the case of non-conditional L-estimates) and we lead a Monte-Carlo comparison study that points out the superiority of the smoothed estimator in finite samples. The proofs of the asymptotic results are given in Appendix A. They are valid under mild β -mixing conditions and can thus be useful in a time series context where outliers are particularly likely to appear as pointed out by Lucas (1996).

2 Robust estimation of regression using L-estimates

2.1 Estimation of conditional cumulative distribution function

Given a $(d + 1)$ -dimensional random vector (\mathbf{X}, Y) , the cumulative distribution function of the random variable Y conditional on the event $\{X = x\}$ is defined by

$$F_{\mathbf{x}}(y) = \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} dv,$$

where $f(\mathbf{x}, y)$ is the joint density of (\mathbf{X}, Y) and $f(\mathbf{x})$ is the marginal density of \mathbf{X} . The common practice in literature (Härdle, 1990) consists in estimating this function using a local empirical conditional distribution function defined by

$$\tilde{F}_x(y) = \sum_{i=1}^n \frac{\mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_x}\right)}{\sum_{j=1}^n \mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_j}{h_x}\right)} I(Y_i \leq y), \quad (1)$$

where I denotes the indicator function and $\mathbf{K}_{\mathbf{x}}$ is a d dimensional kernel. This function, being a local empirical cumulative distribution, has a step function structure.

As pointed out by Fernholz (1997) in the case of non-conditional cumulative distribution function, using an empirical distribution function may not be the best choice for estimating quantiles, and more generally, computing L-estimates. Furthermore, the step structure of the local empirical cumulative distribution function and its weak regularity properties can be difficult to handle in a theoretical framework.

Therefore, we propose to apply an additional smoothing to the variable Y and to estimate $F_{\mathbf{x}}(y)$ by a local smoothed conditional distribution function

$$\hat{F}_{\mathbf{x}}(y) = \sum_{i=1}^n \frac{\mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_x}\right)}{\sum_{j=1}^n \mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_j}{h_x}\right)} K_I\left(\frac{y-Y_i}{h_y}\right) \quad (2)$$

where K_I is a univariate cumulative distribution function (the integral of a kernel). This estimator inherits the regularity properties of the univariate kernel K_I and may thus be used, for instance, for estimating the derivatives of $F_{\mathbf{x}}(y)$.

2.2 Description of the L-empirical and the L-smoothed estimators

Following Boente and Fraiman (1994), we define the conditional L-estimate by

$$m_L(\mathbf{x}) = \int y J\{F_{\mathbf{x}}(y)\} dy \quad (3)$$

where the L-score function J is continuously differentiable with compact support $[a, b] \subset]0; 1[$.

Definition (3) encompasses many useful statistical parameters of interest. If $F_{\mathbf{x}}(y)$ is symmetric around the conditional expectation $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ and if one considers the α -trimming score function $J(u) = (1 - 2\alpha)^{-1} I_{[\alpha; 1-\alpha]}(u)$ with $\alpha \in]0; 1[$, the equality $m_L(\mathbf{x}) = m(\mathbf{x})$ holds. In this case, the α -trimmed conditional expectation $m_L(\mathbf{x})$, can be used to remove outliers and to robustify the estimation of regression. In the limit case $\alpha \rightarrow 1/2$, $m_L(\mathbf{x})$ is equal to the conditional median.

A natural way of estimating $m_L(\mathbf{x})$ consists of plugging an estimator of $F_{\mathbf{x}}(y)$ in expression (3). If one plugs in the empirical conditional cdf (1) as proposed by Boente and Fraiman (1994), $m_L(\mathbf{x})$ is then estimated by the local empirical L-estimator

$$\tilde{m}_L(\mathbf{x}) = \sum_{i=1}^n \frac{\mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_x}\right)}{\sum_{j=1}^n \mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_j}{h_x}\right)} J\left\{\hat{F}_{\mathbf{x}}(Y_i)\right\}. \quad (4)$$

Instead of plugging the empirical conditional cdf, we propose to plug-in the smoothed conditional cdf (2). The function $m_L(\mathbf{x})$ is then estimated by the local smoothed L-estimator

$$\hat{m}_L(\mathbf{x}) = \int y J\left\{\hat{F}_{\mathbf{x}}(y)\right\} dy, \quad (5)$$

whereby the integral is approximated using classical numerical integration routines.

As will be shown in the next section, both the local empirical and the local smoothed L-estimators have the same asymptotic properties. However, we demonstrate that estimator (5) has superior computational and finite sample statistical properties to (4); see Sections 4 and 5, respectively.

3 Asymptotic analysis

We begin by giving asymptotic bounds on the smoothed cumulative distribution function (2).

Lemma 1 *Under assumptions (A1) to (A5) given in Appendix A.1 the following uniform bound holds for $\hat{F}_{\mathbf{x}}(y)$*

$$\sup_{x,y} \left| \hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y) \right| = O_p\left(n^{-1/2}h^{-d} + h^r\right).$$

Proof: See Appendix A.2.

We can now derive the asymptotic distribution of the smoothed L-estimator (5).

Theorem 2 *Under assumptions (A1) to (A5) given in Appendix A.1, the following asymptotic distribution holds for $\hat{m}_L(\mathbf{x})$:*

$$n^{1/2}h^{d/2} \{\hat{m}_L(\mathbf{x}) - m_L(\mathbf{x})\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, V),$$

where

$$V = \int \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} [I(w \leq y) - F_{\mathbf{x}}(y)] dy \right\}^2 f(\mathbf{x}, w) dw \cdot \int K^2(\mathbf{x}) d\mathbf{x}.$$

Proof: See Appendix A.3.

4 Computational comparison

The asymptotic distribution of the smoothed L-estimator $\hat{m}_L(\mathbf{x})$ is identical to that of the empirical L-estimator $\tilde{m}_L(\mathbf{x})$ given by Boente and Fraiman (1994). On the other hand, the computational burden for its computation is much less than for the empirical L-estimator by Boente and Fraiman (1994). To show this, we will compare the computational costs for the L-empirical and the L-smoothed estimators, both from a theoretical and an empirical point of view. We assume that the computation of $\mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_{\mathbf{x}}}\right)$ requires $C_{\mathbf{x}}(d)$ operations, the computation of $K_I\left(\frac{y-Y_i}{h_y}\right)$ requires C_I operations and the computation of the score function J requires C_J operations. Although the cost of computing $\mathbf{K}_{\mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_{\mathbf{x}}}\right)$ depends on the dimension d , we consider d to be fixed and thus $C_{\mathbf{x}}(d)$ to be a constant since d is determined only by the number of employed explanatory variables.

Empirical Estimation The calculation of the empirical cumulative distribution function $\tilde{F}_{\mathbf{x}}(\cdot)$ at one point y using (1) needs $n(C_{\mathbf{x}}(d) + 3)$ operations. Since the computation of $\tilde{m}_L(\mathbf{x})$ requires the computation of $\tilde{F}_{\mathbf{x}}(\cdot)$ at every point Y_i , we

get for the Boente and Fraiman (1994) $\tilde{m}_L(\mathbf{x})$

$$n^2 \{C_{\mathbf{x}}(d) + 3\} + n \{C_J + C_{\mathbf{x}}(d) + 3\} = O(n^2) \quad (6)$$

as the cost of operations.

Smoothed estimation Analogously, the computation of the smoothed cumulative distribution function $\hat{F}_x(\cdot)$ at one point y requires $n(C_{\mathbf{x}}(d) + C_I + 2)$. We assume that the integral in expression (5) is approximated on a grid of k points, whereby k is a fixed constant determined by the required precision of numerical integration. Thus, its computation needs $O(k)$ operations. We do not consider higher order numerical integration methods since the function to integrate is quite regular. Nevertheless, if such methods are used, the results do not change qualitatively. Finally, the computation of $\hat{m}_L(\mathbf{x})$ requires

$$n \{C_{\mathbf{x}}(d) + C_I + 2\} O(k) = O(nk) \quad (7)$$

operations.

Computation time results In order to corroborate our theoretical results, we performed a set of simulations in the univariate case. We used the data generating process

$$Y = m(\mathbf{X}) + \varepsilon, \quad (8)$$

where the regression function is given by

$$m_L(\mathbf{x}) = -1 + \sqrt{x} - x^2$$

and the regressor \mathbf{X} is univariate and uniformly distributed on the interval $[0; 1]$. The error term ε has a normal distribution with standard deviation 0.1. This distribution

Table 1: Computational time in seconds

Sample size n	50	100	200	500	1000	2000
Method (k)						
Empirical	<1	3.6	12.8	77.6	301	1200
Smoothed (25)	<1	1.4	2.0	6.0	11	22
Smoothed (50)	1	2.0	4.6	11.0	22	43
Smoothed (100)	2	4.6	9.0	22.0	43	85
Smoothed (150)	4	6.6	13.4	33.0	65	128
Smoothed (250)	6	11.6	22.0	55.0	107	212
Smoothed (500)	11	22.6	44.2	110	213	425

Table 2: Relative computational time

Sample size n	50	100	200	500	1000	2000
Method (k)						
Empirical	0.4	0.8	1.4	3.5	7.0	14.0
Smoothed (25)	0.3	0.2	0.2	0.3	0.3	0.3
Smoothed (50)	0.5	0.4	0.5	0.5	0.5	0.5
Smoothed (100)	1.0	1.0	1.0	1.0	1.0	1.0
Smoothed (150)	2.0	1.4	1.5	1.5	1.5	1.5
Smoothed (250)	3.0	2.5	2.4	2.5	2.5	2.5
Smoothed (500)	5.5	4.9	4.9	5.0	5.0	5.0

is symmetric so that the regression function $m(\mathbf{X})$ and the α -trimmed expectation $m_L(\mathbf{X})$ are equal. The estimations are performed a 10%-trimming score.

Table 1 and 2 contain the absolute and relative time necessary for the estimation of the empirical estimator (4) and of the smoothed one (5) for different sample sizes and different number of points on the numerical integration grid. Increasing the number of integration points decreases the error of approximation in the integral (4). In our opinion, using 100–200 points is a good choice for applications. Table 1 contains the computation times expressed in seconds, whereas in Table 2 these times are relative to the smoothed estimation using 100 points integration grid. Results in Table 1 confirm our theoretical findings that the smoothed estimator will be faster to compute for large samples. Already when the number of data points is twice the number of points on the integration grid, the smoothed estimator performs better. Furthermore, results in Table 2 support the theoretical conclusions (6) and (7), which imply that these results should be proportional to n for the empirical estimator and constant for the smoothed one.

5 Finite sample comparison

As noticed in Section 3, the empirical L-estimator and the smoothed one need not be compared from an asymptotic point of view since they share the same asymptotic distribution. However, relying on arguments established by Fernholz (1997) in a non-conditional setting, we can create the hypothesis that the smoothed L-estimator has better finite sample properties than the empirical one. At the first glance, this is surprising since the additional smoothing involved in the smoothed conditional cumulative distribution function $\tilde{F}_x(\cdot)$ may cause an additional bias (asymptotically negligible but sensible in finite sample). Nevertheless, as demon-

Table 3: Comparison of the Nadaraya-Watson, the L-empirical and L-smoothed estimators by the mean square error under normal errors. All mean square errors are multiplied by 10^3 .

Method	Empirical	Nadaraya	Smoothed	Smoothed
Grid	$\tilde{m}_L(\mathbf{x})$	Watson	$\hat{m}_L(\mathbf{x})$ ($h_y=0.15$)	$\hat{m}_L(\mathbf{x})$ ($h_y=0.25$)
0.05	1.53	1.51	1.88	1.78
0.10	1.32	0.95	1.00	0.95
0.15	1.11	0.76	0.83	0.75
0.20	0.90	0.61	0.63	0.62
0.25	0.90	0.61	0.65	0.61
0.30	0.76	0.56	0.58	0.57
0.35	0.81	0.59	0.63	0.63
0.40	0.87	0.63	0.67	0.63
0.45	0.83	0.61	0.65	0.64
0.50	0.91	0.64	0.71	0.68
0.55	0.82	0.65	0.74	0.72
0.60	0.71	0.53	0.63	0.58
0.65	0.77	0.54	0.59	0.54
0.70	0.89	0.60	0.62	0.61
0.75	0.98	0.67	0.68	0.70
0.80	1.08	0.74	0.81	0.74
0.85	1.08	0.77	0.84	0.80
0.90	0.99	0.77	0.88	0.85
0.95	1.07	1.28	1.70	1.60

strated by Fernholz (1997), this additional bias goes along with a decrease of the variance of the estimator. Because this decrease of the variance surpasses the additional bias, the smoothing will result in a gain in terms of mean square error of the smoothed conditional L-estimator.

In this section, we attempt to compare the finite sample properties of the empirical and smoothed L-estimators using Monte Carlo simulations. The comparison is made for a range of bandwidth choices and sample sizes as well as for errors coming both from a Gaussian and a heavier-tailed distribution. All simulations are done in the statistical computing environment XploRe.

Table 3 contains the mean square errors for the Nadaraya-Watson, the L-empirical and the L-smoothed estimators (for two different bandwidths h_y) using the data generating process (8). They were calculated at 19 points from 0.05 to 0.95, for 100 observations and a bandwidth $h_x = 0.05$ using 1000 simulations. The L-smoothed estimator clearly outperforms the L-empirical estimator for points ranging from 0.10 to 0.90. At the boundaries, the bias effect of the additional smoothing becomes predominant on the variance decrease so that the L-smoothed estimator performs worse. For comparison, we also computed the Nadaraya-Watson estimator. It has to be noticed that the L-smoothed estimator performs almost as well as the Nadaraya-Watson estimator whereas the L-empirical performs apparently worse. This is particularly interesting since the data contain no outliers and errors are normally distributed. Therefore, using the L-smoothed estimator can be a good strategy for estimating regression even if the presence of outliers in the data is only hypothetical, since the effects of outliers on the classical Nadaraya-Watson estimator can be very damaging.

The sensitivity of the Nadaraya-Watson and robust properties of L-based estimators are documented by Table 4 containing mean square errors for all estimators

Table 4: Comparison of the Nadaraya-Watson, the L-empirical and L-smoothed estimators by the mean square error under t_2 -distributed errors. All mean square errors are multiplied by 10^3 .

Method	Empirical	Nadaraya	Smoothed	Smoothed
Grid	$\tilde{m}_L(\mathbf{x})$	Watson	$\hat{m}_L(\mathbf{x})$ ($h_y=0.15$)	$\hat{m}_L(\mathbf{x})$ ($h_y=0.25$)
0.05	3.98	7.63	7.86	19.61
0.10	1.03	3.44	1.63	5.96
0.15	0.99	4.00	0.97	2.32
0.20	1.15	5.68	1.02	1.67
0.25	1.17	6.66	1.02	1.37
0.30	1.11	5.94	1.02	1.34
0.35	1.06	4.39	0.99	1.48
0.40	1.01	3.33	0.94	1.92
0.45	0.95	2.92	0.93	2.26
0.50	0.95	2.80	0.91	1.97
0.55	0.95	2.71	0.94	1.76
0.60	0.98	2.62	0.92	1.46
0.65	0.95	2.98	0.89	1.26
0.70	1.01	4.13	0.91	1.18
0.75	1.06	6.46	0.99	1.16
0.80	1.15	10.80	1.17	1.86
0.85	1.33	16.72	1.51	5.09
0.90	1.77	21.87	3.47	13.03
0.95	4.04	25.98	11.24	29.37

at 19 points from 0.05 to 0.95. We use again model (8) with 100 observations and bandwidth $h_x = 0.10$, but the error term ε had the Student t_2 distribution with two degrees of freedom this time. Thus, we increase the probability that a large error occurs. The immediate consequence is that the mean square error of all estimates increases. However, the Nadaraya-Watson estimator, which performed best in the simulation using normally distributed data, is now worst of all methods, whereas the L-smoothed and L-empirical estimators are affected much less and are therefore better now. Additionally, we can see that the L-smoothed estimator with y -bandwidth $h_y = 0.15$ still outperforms L-empirical in the central part of the domain, $[0.15, 0.80]$, but its counterpart with $h_y = 0.25$ is significantly worse. Hence, if the L-smoothed estimator is to be robust, we should not oversmooth when estimating the conditional distribution $F_{\mathbf{x}}(y)$.

In order to show that the previous results are not due to a special choice of the bandwidth, we tried different choices of bandwidth for data with normally distributed errors and calculated the average of the mean square errors for points ranging from 0.15 to 0.85. The extreme points of the grid are not considered for the bias reasons discussed earlier. These results are summarized in Table 5 and confirm the previous conclusions.

6 Conclusion

Clearly, our theoretical and empirical results point out the superiority of the L-smoothed estimator over the L-empirical one both for computational and finite sample properties. Although one might argue that with increasing sample size, the difference between the two estimators disappear as suggested by the asymptotic results, it is necessary to keep in mind the high computational burden connected with

Table 5: Comparison of the Nadaraya-Watson, the L-empirical and the L-smoothed estimators for different bandwidths and sample size. All values are multiplied by 10^3 .

Sample size, Bandwidth	Empirical $\tilde{m}_L(\mathbf{x})$	Nadaraya Watson	Smoothed $\hat{m}_L(\mathbf{x})$ ($h_y=0.15$)	Smoothed $\hat{m}_L(\mathbf{x})$ ($h_y=0.25$)
$n = 100, h_x=0.05$	0.90	0.64	0.68	0.65
$n = 100, h_x=0.07$	0.70	0.51	0.50	0.47
$n = 100, h_x=0.10$	0.73	0.58	0.51	0.49
$n = 200, h_x=0.05$	0.41	0.32	0.36	0.33
$n = 200, h_x=0.07$	0.37	0.31	0.29	0.27
$n = 200, h_x=0.10$	0.40	0.37	0.30	0.28

the L-empirical estimator. The use of the smoothed L-estimator is thus indicated in all cases.

A Proofs

A.1 Assumptions

- (A1) : The sequence $\{(\mathbf{X}_i, Y_i)\}_{i=1..n}$ is a sequence of strictly stationary and β mixing realizations of the vector (\mathbf{X}, Y) satisfying $k^\delta \beta_k \rightarrow 0$ for some fixed $\delta > 1$. Here $\beta_k = E \sup \left\{ \left| P(A | \mathcal{F}_m^l) - P(A,) \right| : A \in \mathcal{F}_{k+l}^\infty \right\}$ where $\mathcal{F}_t^{t'}$ is the σ algebra generated by $(\mathbf{X}_t, Y_t), \dots, (\mathbf{X}_{t'}, Y_{t'})$
- (A2) : The density $f(\mathbf{x}, y)$ is compactly supported and admits continuous derivatives up to order r .

- (A3) : The density $f(\mathbf{x})$ admits a strictly positive lower bound, b .
- (A4) : The univariate function K is a symmetric compactly supported kernel of order r .
- (A5) : The bandwidth satisfies $\lim_{n \rightarrow \infty} h = 0$ (the dependence of h on n is left implicit for the simplicity of notations) in such a way that $\lim_{n \rightarrow \infty} n^{1/2} h^{3d/2} \rightarrow \infty$ and $\lim_{n \rightarrow \infty} n^{1/2} h^{r+d/2} \rightarrow 0$.

A.2 Proof of lemma1

$$\begin{aligned}
\left| \hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y) \right| &= \left| \int_{-\infty}^y \frac{\hat{f}(\mathbf{x}, v)}{\hat{f}(\mathbf{x})} dv - \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} dv \right| \\
&\leq \left| \frac{1}{\hat{f}(\mathbf{x})} \right| \left| \int_{-\infty}^y \hat{f}(\mathbf{x}, v) dv - \int_{-\infty}^y f(\mathbf{x}, v) dv \right| \\
&\quad + \left| \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{\hat{f}(\mathbf{x})} dv - \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} dv \right|
\end{aligned}$$

For n large enough, we have almost surely $\sup_{\mathbf{x}} |\hat{f}(\mathbf{x})| \geq \frac{b}{2}$ (which comes from the almost sure uniform convergence of $\hat{f}(\mathbf{x})$ towards $f(\mathbf{x})$ under assumptions (A1) to (A5)). Using $\sup_{\mathbf{x}} |f(\mathbf{x})| \geq b$ (assumption (A3)) we obtain first

$$\sup_{\mathbf{x}, y} \left| \frac{1}{\hat{f}(\mathbf{x})} \right| \left| \int_{-\infty}^y \hat{f}(\mathbf{x}, v) dv - \int_{-\infty}^y f(\mathbf{x}, v) dv \right| \leq 2b^{-1} \sup_{\mathbf{x}, y} \left| \int_{-\infty}^y \hat{f}(\mathbf{x}, v) dv - \int_{-\infty}^y f(\mathbf{x}, v) dv \right|_{\infty}.$$

Using the expression

$$\sup_{\mathbf{x}, y} \left| \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{\hat{f}(\mathbf{x})} dv - \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} dv \right| = \sup_{\mathbf{x}, y} \left\{ \left| \frac{f(\mathbf{x}) - \hat{f}(\mathbf{x})}{\hat{f}(\mathbf{x}) f(\mathbf{x})} \right| \left| \int_{-\infty}^y f(\mathbf{x}, v) dv \right| \right\},$$

we also obtain

$$\sup_{\mathbf{x}, y} \left| \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{\hat{f}(\mathbf{x})} dv - \int_{-\infty}^y \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} dv \right| \leq 2b^{-2} \sup_{\mathbf{x}, v} |f(\mathbf{x}, v)| \sup_{\mathbf{x}} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|$$

so that finally

$$\begin{aligned} \sup_{\mathbf{x}, y} \left| \hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y) \right| &\leq 2b^{-1} \sup_{\mathbf{x}, y} \left| \int_{-\infty}^y \hat{f}(\mathbf{x}, v) dv - \int_{-\infty}^y f(\mathbf{x}, v) dv \right| \\ &\quad + 2b^{-2} \sup_{\mathbf{x}, v} |f(\mathbf{x}, v)| \sup_{\mathbf{x}} \left| \hat{f}(\mathbf{x}) - f(\mathbf{x}) \right| \end{aligned}$$

From Aït-Sahalia (1995), under assumptions (A1) to (A5), we have

$$\sup_{x, y} \left| \int_{-\infty}^y \hat{f}(\mathbf{x}, v) dv - \int_{-\infty}^y f(\mathbf{x}, v) dv \right| = O_p \left(n^{-\frac{1}{2}} h^{-d} + h^r \right)$$

and

$$\sup_x \left| \hat{f}(\mathbf{x}) - f(\mathbf{x}) \right| = O_p \left(n^{-\frac{1}{2}} h^{-d} + h^r \right)$$

so that

$$\sup_{x, y} \left| \hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y) \right| = O_p \left(n^{-\frac{1}{2}} h^{-d} + h^r \right)$$

A.3 Proof of theorem2

Let us denote by T the functional $T(F_{\mathbf{x}}) = \int_{-\infty}^{+\infty} y J\{F_{\mathbf{x}}(y)\} dF_{\mathbf{x}}(y)$ and let's denote by τ the function $\tau(t) = T(F_{\mathbf{x}} + tH_{\mathbf{x}})$ where $H_{\mathbf{x}} : R \rightarrow R$ is a continuously differentiable function with derivative compactly supported satisfying $\sup_{\mathbf{x}, y} |H_{\mathbf{x}}(y)| < \infty$.

We have

$$\begin{aligned} \tau'(t) &= \int_{-\infty}^{+\infty} y H_{\mathbf{x}}(y) J'\{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} d\{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} \\ &\quad + \int_{-\infty}^{+\infty} y J(F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)) dH_{\mathbf{x}}(y). \end{aligned}$$

An integration by parts gives us

$$\begin{aligned} \int_{-\infty}^{+\infty} y H_{\mathbf{x}}(y) J'\{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} d(F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)) &= [y H_{\mathbf{x}}(y) J\{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\}]_{+\infty}^{-\infty} \\ &\quad - \int_{-\infty}^{+\infty} H_{\mathbf{x}}(y) J\{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} dy \\ &\quad - \int_{-\infty}^{+\infty} y J\{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} dH_{\mathbf{x}}(y) \end{aligned}$$

so that

$$\tau'(t) = - \int_{-\infty}^{+\infty} H_{\mathbf{x}}(y) J \{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} dy.$$

In particular, for $t = 0$, we obtain

$$\tau'(0) = - \int_{-\infty}^{+\infty} H_{\mathbf{x}}(y) J \{F_{\mathbf{x}}(y)\} dy$$

The second derivative of τ is

$$\tau''(t) = - \int_{-\infty}^{+\infty} H_{\mathbf{x}}^2(y) J' \{F_{\mathbf{x}}(y) + tH_{\mathbf{x}}(y)\} dy$$

so that, under assumptions (A2), we have, for all $t \in [0; 1]$

$$|\tau''(t)| = O \left(\sup_{\mathbf{x}, y} |H_{\mathbf{x}}(y)|^2 \right).$$

Now a Taylor expansion of τ between 0 and 1 gives us :

$$T(F_{\mathbf{x}} + H_{\mathbf{x}}) = T(F_{\mathbf{x}}) - \int_{-\infty}^{+\infty} H_{\mathbf{x}}(y) J \{F_{\mathbf{x}}(y)\} dy + O \left(\sup_{\mathbf{x}, y} |H_{\mathbf{x}}(y)|^2 \right) \quad (9)$$

Taking $H_{\mathbf{x}}(y) = \hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y)$ in expression (9) and using Lemma 1, we obtain

$$\tilde{m}(\mathbf{x}) - m(x) = - \int_{-\infty}^{+\infty} [\hat{F}_{\mathbf{x}}(y) - F_{\mathbf{x}}(y)] J \{F_{\mathbf{x}}(y)\} dy + O_p(n^{-1}h^{-2d} + h^{2r})$$

Let us study the leading order term

$$L_n = - \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^y \frac{\hat{f}(\mathbf{x}, v)}{\hat{f}(\mathbf{x})} - \frac{f(\mathbf{x}, v)}{f(\mathbf{x})} dv \right\} J \{F_{\mathbf{x}}(y)\} dy$$

$$L_n = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^y \frac{(\hat{f}(\mathbf{x}, v) - f(\mathbf{x}, v)) f(v) - f(\mathbf{x}, v) (\hat{f}(\mathbf{x}) - f(\mathbf{x}))}{\hat{f}(\mathbf{x}) f(\mathbf{x})} dv \right] J \{F_{\mathbf{x}}(y)\} dy$$

using $\sup_x \left| \frac{\hat{f}(\mathbf{x}) - f(\mathbf{x})}{\hat{f}(\mathbf{x})} \right| = o_p(1)$ (which holds under assumptions (A1) to (A5)). We

get

$$L_n = \left\{ \int_{-\infty}^{+\infty} \left[\int_{-\infty}^y \frac{(\hat{f}(\mathbf{x}, v) - f(\mathbf{x}, v)) f(\mathbf{x}) - f(\mathbf{x}, v) (\hat{f}(\mathbf{x}) - f(\mathbf{x}))}{f^2(\mathbf{x})} dv \right] J \{F_{\mathbf{x}}(y)\} dy \right\} (1 + o_p(1))$$

so that, using Slutsky theorem, we only have to study the asymptotic distribution of

$$\tilde{L}_n = \left\{ \int_{-\infty}^{+\infty} \left[\int_{-\infty}^y \frac{(\hat{f}(\mathbf{x}, v) - f(\mathbf{x}, v)) f(\mathbf{x}) - f(\mathbf{x}, v) (\hat{f}(\mathbf{x}) - f(\mathbf{x}))}{f^2(\mathbf{x})} dv \right] J\{F_{\mathbf{x}}(y)\} dy \right\}$$

If we define $\mathbf{K}_h(\cdot) = \frac{1}{h} \mathbf{K}\left(\frac{\cdot}{h}\right)$, we have

$$\tilde{L}_n = \frac{1}{n} \sum_{i=1}^n \int \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \left[\begin{array}{c} \mathbf{K}_h(\mathbf{x} - \mathbf{X}_i) \int_{-\infty}^y K_h(v - Y_i) dv \\ - \int_{-\infty}^y f(\mathbf{x}, v) dv - F_{\mathbf{x}}(y) (K_h(\mathbf{x} - \mathbf{X}_i) - f(\mathbf{x})) \end{array} \right] dy \right\}.$$

We are going to separate the study of \tilde{L}_n into a determinist ‘bias’ term \tilde{B}_n and a stochastic ‘variance’ term \tilde{V}_n :

$$\tilde{B}_n = \int \left\{ \int_{-\infty}^{+\infty} \left[\frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \mathbf{K}_h(\mathbf{x} - \mathbf{u}) \left(\int_{-\infty}^y K_h(v - w) dv - F_{\mathbf{x}}(y) \right) \right] dy \right\} f(\mathbf{u}, w) d\mathbf{u} dw.$$

With the change of variable $\boldsymbol{\xi} = \frac{\mathbf{x} - \mathbf{u}}{h}$ in the integration with respect to \mathbf{u} and $\varsigma = \frac{y - w}{h}$ in the integration with respect to w , we obtain

$$\tilde{B}_n = \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \left[\int_{-\infty}^y \int \int \mathbf{K}(\boldsymbol{\xi}) (K(\varsigma) - F_{\mathbf{x}}(y)) f(\mathbf{x} - h\boldsymbol{\xi}, v - h\varsigma) d\boldsymbol{\xi} d\varsigma dv \right] dy.$$

Using a Taylor expansion of $f(\mathbf{x} - h\boldsymbol{\xi}, v - h\varsigma)$ and of $f(\mathbf{x} - h\boldsymbol{\xi})$ up to order r , under assumption (A4), we obtain $\tilde{B}_{2n} = O(h^r)$ and, under assumption (A5), we have $n^{1/2} h^{d/2} \tilde{B}_{2n} = o_p(1)$.

We now have to study the ‘variance’ term normalized at rate $n^{1/2} h^{d/2}$

$$n^{1/2} h^{d/2} \tilde{V}_n = \sum_{i=1}^n \sqrt{\frac{h^d}{n}} \left[\int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \mathbf{K}_h(\mathbf{x} - \mathbf{X}_i) \left(\int_{-\infty}^y K_h(v - Y_i) dv - F_{\mathbf{x}}(y) \right) dy - \tilde{B}_{2n} \right]$$

Let us define $\mathcal{F}_{n,i}$ the σ field generated by $\{\mathbf{X}_j, Y_j\}_{j=1, \dots, i}$ and let us define the random variable

$$Z_{n,i} = \sqrt{\frac{h^d}{n}} \left[\int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \mathbf{K}_h(\mathbf{x} - \mathbf{X}_i) \left(\int_{-\infty}^y K_h(v - Y_i) dv - F_{\mathbf{x}}(y) \right) dy - \tilde{B}_{2n} \right]$$

such as $n^{1/2}h^{d/2}\tilde{V}_n = \sum_{j=1}^n Z_{n,j}$. The array $\left\{ \sum_{j=1}^i Z_{n,j}, \mathcal{F}_{n,i}, 1 \leq i \leq n, n \geq 1 \right\}$ is a zero mean, square integrable martingale array. We have

$$\sum_{j=1}^i E(Z_{n,j}^2) = h^d \int \int \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} K_h(\mathbf{x} - \mathbf{u}) \left[\int_{-\infty}^y K_h(v - w) dv - F_{\mathbf{x}}(y) \right] dy \right\}^2 f(\mathbf{u}, w) d\mathbf{u} dw - h^d (\tilde{B}_{2n})^2$$

With the change of variable $\boldsymbol{\xi} = \frac{\mathbf{x} - \mathbf{u}}{h}$ in the integration with respect to \mathbf{u} , we obtain :

$$\sum_{j=1}^i E(Z_{n,j}^2) = \int \left\langle \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} \left[\int_{-\infty}^y K_h(v - w) dv - F_{\mathbf{x}}(y) \right] dy \right\}^2 \cdot \right\rangle dw - h^d (\tilde{B}_{2n})^2 \cdot \left\{ \int K^2(\boldsymbol{\xi}) f(\mathbf{x} - h\boldsymbol{\xi}, w) d\boldsymbol{\xi} \right\}$$

With $\lim_{n \rightarrow \infty} \int_{-\infty}^y K_h(v - w) dv = 1$ if $y \geq w$ and 0 otherwise, and with $\tilde{B}_{2n} = O(h^r)$, we obtain

$$\lim_{n \rightarrow \infty} E \left(\sum_{j=1}^n Z_{n,j}^2 \right) = \int \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} [I_{]-\infty, y]}(w) - F_{\mathbf{x}}(y) \right\}^2 dy \cdot f(\mathbf{x}, w) dw \cdot \int K^2(\boldsymbol{\xi}) d\boldsymbol{\xi}$$

so that

$$\sum_{j=1}^n E(Z_{n,j}^2 / \mathcal{F}_{n,j-1}) \xrightarrow{P} \int \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} [I_{]-\infty, y]}(w) - F_{\mathbf{x}}(y) \right\}^2 dy \cdot f(\mathbf{x}, w) dw \cdot \int K^2(\boldsymbol{\xi}) d\boldsymbol{\xi}.$$

Under assumptions (A1) to (A4), conditions 3.19 and 3.20 of Corollary 3.1 of Hall and Heyde (1981, pp 58) are satisfied, so that we obtain

$$n^{1/2}h^{d/2}\tilde{V}_n \rightarrow \mathcal{N}(0, V),$$

where

$$V = \int \left\{ \int_{-\infty}^{+\infty} \frac{J\{F_{\mathbf{x}}(y)\}}{f(\mathbf{x})} [I_{]-\infty, y]}(w) - F_{\mathbf{x}}(y) \right\}^2 dy \cdot f(\mathbf{x}, w) dw \cdot \int K^2(\boldsymbol{\xi}) d\boldsymbol{\xi}.$$

References

- [1] Y. Aït-Sahalia, The delta method for nonlinear kernel functionals, *Ph. D. dissertation*, University of Chicago, USA, 1995.
- [2] V. Barnett and T. Lewis, *Outliers in statistical data*, Wiley, New York, 1979.
- [3] G. Boente and R. Fraiman, Local L-estimators for nonparametric regression under dependence, *Journal of nonparametric statistics* **4**, (1994), 91–101.
- [4] L. T. Fernholz, Reducing the variance by smoothing, *Journal of Statistical Planning and Inference*, **57**, (1997), 29–38.
- [5] P. Hall and C. C. Heyde, *Martingale Limit Theory and its Applications*, Academic Press, New York, 1981.
- [6] W. Härdle, *Applied nonparametric regression*, Cambridge University Press, Cambridge, 1990.
- [7] A. Lucas, Outlier robust unit root analysis, *Ph.D. dissertation*, Tinbergen Institute, Rotterdam, 1996.
- [8] E. A. Nadaraya, On estimating regression, *Theory. Prob. App*, **10**, (1964), 186–190.
- [9] G. S. Watson, Smooth regression analysis, *Sankhya*, Ser. A, **26**, (1964), 359–372.